Supplement to “Familywise Error Rate Control by Interactive Unmasking”

Boyan Duan [1] **Aaditya Ramdas** [1 2] **Larry Wasserman** [1 2]

1. Distribution of the null p-values

Error control holds for null p-values whose distribution satisfies a property called *mirror-conservativeness*:

$$f(a) \leq f\left(1 - \frac{1-p_*}{p_*}a\right), \quad \text{for all } 0 \leq a \leq p_*, \tag{1}$$

where f is the probability mass function of P for discrete p-values or the density function otherwise, and p_* is the parameter in Algorithm 1. The mirror-conservativeness is first proposed by Lei & Fithian (2018) in the case of $p_* = 0.5$. A more commonly used notion of conservativeness is that p-values are stochastically larger than uniform:

$$\mathbb{P}(P \leq a) \leq a, \quad \text{for all } 0 \leq a \leq 1,$$

which neither implies nor is implied by the mirror-conservativeness.

Sufficient conditions of the mirror-conservativeness include that f is nondecreasing or the CDF of the p-value is convex. For example, consider a one-dimensional exponential family and the hypotheses to test the value of its parameter θ:

$$H_0 : \theta \leq \theta_0, \quad \text{versus} \quad H_1 : \theta > \theta_0,$$

where θ_0 is a prespecified constant. The p-value calculated from the uniformly most powerful test is shown to have a nondecreasing density (Zhao et al., 2019); thus, it satisfies the mirror-conservativeness. The conservative nulls described in Section 4.1 also fall into the above category where the exponential family is Gaussian, and the parameter is the mean value. In this setting, the i-FWER test has a valid error control as proved in Appendix 2 (for the tent masking) and Appendix 6 (for other masking functions).

2. Proof of Theorem 1

The main idea of the proof is that the missing bits $h(P_i)$ of nulls are coin flips with probability p_* to be heads, so the number of false rejections (i.e. the number of nulls with $h(P_i) = 1$ before the number of hypotheses with $h(P_i) = -1$ reaches a fixed number) is stochastically dominated by a negative binomial distribution. There are two main challenges. First, the interaction uses unmasked p-value information to reorder $h(P_i)$, so it is not trivial to show that the reordered $h(P_i)$ preserve the same distribution as that before ordering. Second, our procedure runs backward to find the first time that the number of hypotheses with negative $h(P_i)$ is below a fixed number, which differs from the standard description of a negative binomial distribution.

2.1. Missing bits after interactive ordering

We first study the effect of interaction. Imagine that Algorithm 1 does not have a stopping rule and generates a full sequence of $\mathcal{R}_t$ for $t = 0, 1, \ldots n$, where $\mathcal{R}_0 = [n]$ and $\mathcal{R}_n = \emptyset$. It leads to an ordered sequence of $h(P_i)$:

$$h(P_{\pi_1}), h(P_{\pi_2}), \ldots, h(P_{\pi_n}),$$

where π_n is the index of the first excluded hypothesis and π_j denotes the index of the hypothesis excluded at step $n - j + 1$, that is $\pi_j = \mathcal{R}_{n-j} \backslash \mathcal{R}_{n-j+1}$.

[1]Department of Statistics and Data Science, Carnegie Mellon University [2]Machine Learning Department, Carnegie Mellon University. Correspondence to: Boyan Duan <boyand@stat.cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

Lemma 1. *Suppose the null p-values are uniformly distributed and all the hypotheses are nulls, then for any $j = 1, \dots, n$,*

$$\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] = p_*,$$

and $\{\mathbb{1}\left(h(P_{\pi_j}) = 1\right)\}_{j=1}^n$ are mutually independent.

Proof. Recall that the available information for the analyst to choose π_j is $\mathcal{F}_{n-j} = \sigma\Big(\{x_i, g(P_i)\}_{i=1}^n, \{P_i\}_{i \notin \mathcal{R}_{n-j}}\Big)$. First, consider the conditional expectation:

$$\begin{aligned}
& \mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \mathcal{F}_{n-j}\right] \\
= & \sum_{i \in [n]} \mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \pi_j = i, \mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i | \mathcal{F}_{n-j}\right) \\
\overset{(a)}{=} & \sum_{i \in \mathcal{R}_{n-j}} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right) | \pi_j = i, \mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i | \mathcal{F}_{n-j}\right) \\
\overset{(b)}{=} & \sum_{i \in \mathcal{R}_{n-j}} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right) | \mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i | \mathcal{F}_{n-j}\right) \\
\overset{(c)}{=} & \sum_{i \in \mathcal{R}_{n-j}} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right)\right] \mathbb{P}\left(\pi_j = i | \mathcal{F}_{n-j}\right) \\
= & \; p_* \sum_{i \in \mathcal{R}_{n-j}} \mathbb{P}\left(\pi_j = i | \mathcal{F}_{n-j}\right) = p_*,
\end{aligned} \tag{2}$$

where equation (a) narrows down the choice of i because $\mathbb{P}(\pi_j = i \mid \mathcal{F}_{n-j}) = 0$ for any $i \notin \mathcal{R}_{n-j}$; equation (b) drops the condition of $\pi_j = i$ because π_j is measurable with respect to $\mathcal{F}_{n-j}$; and equation (c) drops the condition $\mathcal{F}_{n-j}$ because by the independence assumptions in Theorem 1, $h(P_i)$ is independent of $\mathcal{F}_{n-j}$ for any $i \in \mathcal{R}_{n-j}$.

Therefore, by the law of iterated expectations, we prove the claim on expected value:

$$\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \mathcal{F}_{n-j}\right]\right] = p_*.$$

For mutual independence, we can show that for any $1 \leq k < j \leq n$, $\mathbb{1}\left(h(P_{\pi_k}) = 1\right)$ is independent of $\mathbb{1}\left(h(P_{\pi_j}) = 1\right)$. Consider the conditional expectation:

$$\begin{aligned}
& \mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_k}) = 1\right) \middle| \mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] \\
= & \; \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_k}) = 1\right) \middle| \mathcal{F}_{n-k}, \mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] \middle| \mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] \\
& \text{(note that } \mathbb{1}\left(h(P_{\pi_j}) = 1\right) \text{ is measurable with respect to } \mathcal{F}_{n-k}) \\
= & \; \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_k}) = 1\right) | \mathcal{F}_{n-k}\right] \middle| \mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] \\
& \text{(use equation (2) for the conditional expectation)} \\
= & \; \mathbb{E}\left[p_* \mid \mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] = p_*.
\end{aligned}$$

It follows that $\mathbb{1}\left(h(P_{\pi_k}) = 1\right) \mid \mathbb{1}\left(h(P_{\pi_j}) = 1\right)$ is a Bernoulli with parameter p_*, same as the marginal distribution of $\mathbb{1}\left(h(P_{\pi_k}) = 1\right)$; thus, $\mathbb{1}\left(h(P_{\pi_k}) = 1\right)$ is independent of $\mathbb{1}\left(h(P_{\pi_j}) = 1\right)$ for any $1 \leq k < j \leq n$ as stated in the Lemma. $\square$

Corollary 1. *Suppose the null p-values are uniformly distributed and there may exist non-nulls. For any $j = 1, \dots, n$,*

$$\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \left\{\mathbb{1}\left(h(P_{\pi_k}) = 1\right)\right\}_{k=j+1}^n, \left\{\mathbb{1}\left(\pi_k \in \mathcal{H}_0\right)\right\}_{k=j+1}^n, \pi_j \in \mathcal{H}_0\right] = p_*,$$

where $\{\pi_k\}_{k=j+1}^n$ represents the hypotheses excluded before π_j.

Proof. Denote the condition $\sigma\left(\{\mathbb{1}\left(h(P_{\pi_k})=1\right)\}_{k=j+1}^{n}, \{\mathbb{1}\left(\pi_k \in \mathcal{H}_0\right)\}_{k=j+1}^{n}\right)$ as $\mathcal{F}_{n-j}^{h}$. The proof is similar to Lemma 1. First, consider the expectation conditional on $\mathcal{F}_{n-j}$:

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \\
=\ &\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \quad \text{(since } \mathcal{F}_{n-j}^{h} \text{ is a subset of } \mathcal{F}_{n-j}\text{)} \\
=\ &\sum_{i\in[n]} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right) \mid \pi_j = i, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) \\
=\ &\sum_{i\in\mathcal{R}_{n-j}\cap\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right) \mid \pi_j = i, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) \\
=\ &\sum_{i\in\mathcal{R}_{n-j}\cap\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right) | \mathcal{F}_{n-j}\right] \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) \\
=\ &p_* \sum_{i\in\mathcal{R}_{n-j}\cap\mathcal{H}_0} \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) = p_*,
\end{aligned} \tag{3}
$$

where we use the same technics of proving equation (2).

Thus, by the law of iterated expectations, we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0\right] \\
=\ &\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right]\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0\right] = p_*,
\end{aligned}
$$

which completes the proof. □

Corollary 2. *Suppose the null p-values can be mirror-conservative as defined in (1) and there may exist non-nulls, then for any $j = 1, \ldots, n$,*

$$
\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\{\mathbb{1}\left(h(P_{\pi_k})=1\right)\}_{k=j+1}^{n}, \{\mathbb{1}\left(\pi_k \in \mathcal{H}_0\right)\}_{k=j+1}^{n}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right] \le p_*,
$$

where $\{g(P_{\pi_k})\}_{k=1}^{n}$ denotes $g(P)$ for all the hypotheses (excluded or not).

Proof. First, we claim that a mirror-conservative p-value P satisfies that

$$
\mathbb{E}\left[\mathbb{1}\left(h(P)=1\right) \mid g(P)\right] \le p_*, \tag{4}
$$

since for every $a \in (0, p_*)$,

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P)=1\right) \mid g(P) = a\right] \\
=\ &\frac{p_* f(a)}{p_* f(a) + (1-p_*) f\left(1 - \frac{1-p_*}{p_*} a\right)} \\
=\ &\frac{p_*}{p_* + (1-p_*) f\left(1 - \frac{1-p_*}{p_*} a\right) / f(a)} \le p_*,
\end{aligned}
$$

where recall that f is the probability mass function of P for discrete p-values or the density function otherwise. The last inequality comes from the definition of mirror-conservativeness in (1). The rest of the proof is similar to Corollary 1, where we first condition on $\mathcal{F}_{n-j}$:

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right] \\
=\ &\sum_{i\in\mathcal{R}_{n-i}\cap\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right) \mid \mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i\middle|\mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right) \\
\overset{(a)}{=}\ &\sum_{i\in\mathcal{R}_{n-i}\cap\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right) \mid g(P_i)\right] \mathbb{P}\left(\pi_j = i\middle|\mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right) \\
\le\ &p_* \sum_{i\in\mathcal{R}_{n-i}\cap\mathcal{H}_0} \mathbb{P}\left(\pi_j = i\middle|\mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right) = p_*,
\end{aligned}
$$

where equation (a) simplify the condition of $\mathcal{F}_{n-j}$ to $g(P_i)$ because for any $i \in \mathcal{R}_{n-i} \cap \mathcal{H}_0$, $h(P_i)$ is independent of other information in $\mathcal{F}_{n-j}$.

Then, by the law of iterated expectations, we obtain

$$\begin{aligned} & \mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \mathcal{F}^h_{n-j}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^n\right] \\ = \; & \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \mathcal{F}_{n-j}, \mathcal{F}^h_{n-j}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^n\right] \middle| \mathcal{F}^h_{n-j}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^n\right] \le p_*, \end{aligned}$$

thus the proof is completed. □

2.2. Negative binomial distribution

In this section, we discuss several procedures for Bernoulli trials (coin flips) and their connections with the negative binomial distribution.

Lemma 2. *Suppose $A_1, \ldots, A_n$ are i.i.d. Bernoulli with parameter p_*. For $t = 1, \ldots, n$, consider the sum $M_t = \sum_{j=1}^t A_j$ and the filtration $\mathcal{G}^o_t = \sigma\left(\{A_j\}_{j=1}^t\right)$. Define a stopping time parameterized by a constant $v (\ge 1)$:*

$$\tau^o = \min\{0 < t \le n : t - M_t \ge v \text{ or } t = n\}, \tag{5}$$

then M_{τ^o} is stochastically dominated by a negative binomial distribution:

$$M_{\tau^o} \preceq \mathrm{NB}(v, p_*).$$

Proof. Recall that the negative binomial $\mathrm{NB}(v, p_*)$ is the distribution of the number of success in a sequence of independent and identically distributed Bernoulli trials with probability p_* before a predefined number v of failures have occurred. Imagine the sequence of A_j is extended to infinitely many Bernoulli trials: $A_1, \ldots, A_n, A'_{n+1}, \ldots$, where $\{A'_j\}_{j=n+1}^\infty$ are also i.i.d. Bernoulli with parameter p_* and they are independent of $\{A_j\}_{j=1}^n$. Let U be the number of success before v-th failure, then by definition, U follows a negative binomial distribution $\mathrm{NB}(v, p_*)$. We can rewrite U as a sum at a stopping time: $U \equiv M_{\tau'}$, where $\tau' = \min\{t > 0 : t - M_t \ge v\}$. By definition, $\tau^o \le \tau'$ (a.s.), which indicates $M_{\tau^o} \le M_{\tau'}$ because M_t is nondecreasing with respect to t. Thus, we have proved that $M_{\tau^o} \preceq \mathrm{NB}(v, p_*)$. □

Corollary 3. *Following the setting in Lemma 2, we consider the shrinking sum $\widetilde{M}_t = \sum_{j=1}^{n-t} A_j$ for $t = 0, 1, \ldots, n-1$. Let the filtration be $\widetilde{\mathcal{G}}_t = \sigma\left(\widetilde{M}_t, \{A_j\}_{j=n-t+1}^n\right)$. Given a constant $v (\ge 1)$, we define a stopping time:*

$$\widetilde{\tau} = \min\{0 \le t < n : (n-t) - \widetilde{M}_t < v \text{ or } t = n-1\}, \tag{6}$$

then it still holds that $\widetilde{M_{\widetilde{\tau}}} \preceq \mathrm{NB}(v, p_*)$.

Proof. We first replace the notion of time t by $n - s$, and let time runs backward: $s = n, n-1, \ldots, 1$. The above setting can be rewritten as $\widetilde{M}_t (= \sum_{j=1}^{n-t} A_j) \equiv M_{n-t} \equiv M_s$ and $\widetilde{\mathcal{G}}_t = \sigma\left(M_s, \{A_j\}_{j=s+1}^n\right) =: \mathcal{G}^b_s$. Define a stopping time:

$$\tau^b = \max\{0 < s \le n : s - M_s < v \text{ or } s = 1\}, \tag{7}$$

which runs backward with respect to the filtration $\mathcal{G}^b_s$. By definition, we have $n - \widetilde{\tau} \equiv \tau^b$, and hence $\widetilde{M_{\widetilde{\tau}}} \equiv M_{\tau^b}$.

Now, we show that $M_{\tau^b} \equiv M_{\tau^o}$ for τ^o defined in Lemma 2. First, consider two edge cases: (1) if $t - M_t < v$ holds for every $0 < t \le n$, then $\tau^b = n = \tau^o$, and thus $M_{\tau^b} = M_{\tau^o}$; (2) if $t - M_t \ge v$ holds for every $0 < t \le n$, then $\tau^b = 1 = \tau^o$, and again $M_{\tau^b} = M_{\tau^o}$. Next, consider the case where $t - M_t < v$ for some t, and $t - M_t \ge v$ for some other t. Note that by definition, $\tau^b + 1$ is a stopping time with respect to $\mathcal{G}^o_t$, and $\tau^b + 1 = \tau^o$. Also, note that by the definition of τ^o, we have $A_{\tau^o} = 0$, so $M_{\tau^o - 1} = M_{\tau^o}$. Thus, $M_{\tau^b} = M_{\tau^o - 1} = M_{\tau^o}$. Therefore, by Lemma 2, $\widetilde{M_{\widetilde{\tau}}} \equiv M_{\tau^b} \equiv M_{\tau^o} \preceq NB(v, p_*)$, as stated in the above Corollary. □

Corollary 4. *Consider a weighted version of the setting in Corollary 3. Let the weights $\{W_j\}_{j=1}^n$ be a sequence of Bernoulli, such that (a) $\sum_{j=1}^n W_j = m$ for a fixed constant $m \le n$; and (b) $A_j \mid \sigma\left(\{A_k, W_k\}_{k=j+1}^n, W_j = 1\right)$ is a Bernoulli with* *parameter p_*. Consider the sum $M_t^w = \sum_{j=1}^{n-t} W_j A_j$. Given a constant $v (\geq 1)$, we define a stopping time:*

$$\begin{aligned} \tau^w = & \min\{0 \leq t < n : \sum_{j=1}^{n-t} W_j(1 - A_j) < v \textit{ or } t = n-1\} \\ \equiv & \min\{0 \leq t < n : \sum_{j=1}^{n-t} W_j - M_t^w < v \textit{ or } t = n-1\}, \end{aligned} \tag{8}$$

then it still holds that $M_{\tau^w}^w \preceq \mathrm{NB}(v, p_*)$.

Proof. Intuitively, adding the binary weights should not change the distribution of the sum $M_{\tau^w}^w = \sum_{j=1}^{n-\tau^w} W_j A_j$, since by condition (b), A_j is still a Bernoulli with parameter p_* when it is counted in the sum. We formalize this idea as follows.

Let $\{B_l\}_{l=1}^m$ be a sequence of i.i.d. Bernoulli with parameter p_*, and denote the sum $\sum_{l=1}^{m-s} B_l$ as $\widetilde{M_s}(B)$. Let $T(t) = m - \sum_{j=1}^{n-t} W_j$, then the stopping time τ^w can be rewritten as

$$\tau^w \equiv \min\{0 \leq t < n : m - T(t) - \widetilde{M_{T(t)}}(B) < v \text{ or } t = n-1\}, \tag{9}$$

because $m - T(t) = \sum_{j=1}^{n-t} W_j$ by definition, and

$$\widetilde{M_{T(t)}}(B) = \sum_{l=1}^{m-T(t)} B_l \overset{d}{=} \sum_{j=1}^{n-t} W_j A_j = M_t^w. \tag{10}$$

For simple notation, we present the reasoning of equation (10) when $t = 0$ (for arbitrary t, consider the distributions conditional on $\{A_k, W_k\}_{k=n-t+1}^n$). That is, we show that $\mathbb{P}(\sum_{l=1}^m B_l = x) = \mathbb{P}(\sum_{j=1}^n W_j A_j = x)$ for every $x \geq 0$. Let $\{b_l\}_{j=1}^m \in \{0,1\}^m$, then we derive that

$$\mathbb{P}(\sum_{l=1}^m B_l = x) = \sum_{\sum b_l = x} \mathbb{P}(B_l = b_l \text{ for } l = 1, \ldots, n) = \sum_{\sum b_l = x} \prod_{l=1}^m f^B(b_l),$$

where f^B is the probability mass function of a Bernoulli with parameter p_*. Let $\{a_k\}_{k=1}^{n-m} \in \{0,1\}^{n-m}$, then for the weighted sum,

$$\begin{aligned} & \mathbb{P}(\sum_{j=1}^n W_j A_j = x) \\ = & \sum_{\sum b_l = x} \sum_{\sum w_j = m} \sum_{a_k} \mathbb{P}(A_j = b_l \text{ if } w_j = 1; A_j = a_k \text{ if } w_j = 0; W_j = w_j \text{ for } i = 1, \ldots, n) \\ = & \sum_{\sum b_l = x} \prod_{l=1}^m f^B(b_l) \underbrace{\sum_{\sum w_j = m} \sum_{\sum a_k} \prod_{w_j = 0} \mathbb{P}(A_j = a_k \mid \sigma\left(\{A_k, W_k\}_{k=j+1}^n, W_j = 0\right) \prod_{j=1}^n \mathbb{P}(W_j = w_j \mid \{A_k, W_k\}_{k=j+1}^n)}_{C \quad \text{(a constant with respect to } x\text{)}} \\ = & \; C \sum_{\sum b_l = x} \prod_{l=1}^m f^B(b_l) = C\mathbb{P}(\sum_{l=1}^m B_l = x), \end{aligned}$$

for every possible value $x \geq 0$, which implies that $\mathbb{P}(\sum_{l=1}^m B_l = x)$ and $\mathbb{P}(\sum_{j=1}^n W_j A_j = x)$ have the same value; and hence we conclude equation (10). It follows that the filtration for both the stopping time τ^w and the sum $M_{t^w}^w$, denoted as $\sigma\left(\sum_{j=1}^{n-t} W_j, M_{t^w}^w, \{A_j, W_j\}_{j=n-t+1}^n\right)$, has the same probability measure as $\sigma\left(m - T(t), \widetilde{M_{T(t)}}(B), \{A_j, W_j\}_{j=n-t+1}^n\right)$. Thus, the sums at the stopping time have the same distribution, $M_{\tau^w}^w \overset{d}{=} \widetilde{M_{T(\tau^w)}}(B)$. The proof completes if $\widetilde{M_{T(\tau^w)}}(B) \preceq \mathrm{NB}(v, p_*)$. It can be proved once noticing that stopping rule (9) is similar to stopping rule (6) except $T(t)$ is random because of W_j, so we can condition on $\{W_j\}_{j=1}^n$ and apply Corollary 3; and this concludes the proof. $\square$

Corollary 5. *In Corollary 4, consider A_j with different parameters. Suppose $A_j \mid \sigma\left(\{A_k, W_k\}_{k=j+1}^n, W_j = 1\right)$ is a Bernoulli with parameter $p\left(\{A_k, W_k\}_{k=j+1}^n\right)$ for every $j = 1, \ldots, n$. Given a constant $p_* \in (0,1)$, if the parameters satisfy that $p\left(\{A_k, W_k\}_{k=j+1}^n\right) \le p_*$ for all $j = 1, \ldots, n$, then it still holds that $M_{\tau^w}^w \preceq \mathrm{NB}(v, p_*)$.*

Proof. We first construct Bernoulli with parameter p_* based on A_j by an iterative process. Start with $j = n$. Let C_n be a Bernoulli independent of $\{A_k\}_{k=1}^n$ with parameter $\frac{p_* - p_n}{1 - p_n}$, where $p_n = \mathbb{E}(A_n \mid W_n = 1)$. Construct

$$B_n = A_n \mathbb{1}(A_n = 1) + C_n \mathbb{1}(A_n = 0), \tag{11}$$

which thus satisfies that $\mathbb{E}(B_n \mid W_n = 1) = p_*$, and that $B_n \ge A_n$ (a.s.). Now, let $j = j - 1$ where we consider the previous random variable. Let C_j be a Bernoulli independent of $\{A_k\}_{k=1}^j$, with parameter

$$\frac{p_* - \widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right)}{1 - \widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right)}, \tag{12}$$

where $\widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right) = \mathbb{E}\left[A_j \mid \sigma\left(\{B_k, W_k\}_{k=j+1}^n, W_j = 1\right)\right]$ (note that the parameter for C_j is well-defined since $\widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right) \le p_*$ by considering the expectation further conditioning on $\{A_k\}_{k=j+1}^n$). Then, we construct B_j as

$$B_j = A_j \mathbb{1}(A_j = 1) + C_j \mathbb{1}(A_j = 0), \tag{13}$$

which thus satisfies that $\mathbb{E}\left[B_j \mid \sigma\left(\{B_k, W_k\}_{k=j+1}^n, W_j = 1\right)\right] = p_*$, and that $B_j \ge A_j$ (a.s.).

Now, consider two procedures for $\{A_j\}_{j=1}^n$ and $\{B_j\}_{j=1}^n$ with the same stopping rule (8) in Corollary 4, where the sum of A_j is denoted as $M_t^w(A)$ and the stopping time as τ_A^w (and the similar notation for B_j). Since construction (13) ensures that $B_j \ge A_j$ for every $j = 1, \ldots, n$, we have $M_t^w(B) \ge M_t^w(A)$ for every t; and hence, $\tau_A^w \ge \tau_B^w$. It follows that

$$M_{\tau_A^w}^w(A) \le M_{\tau_B^w}^w(A) \le M_{\tau_B^w}^w(B) \preceq \mathrm{NB}(v, p_*),$$

where the first inequality is because M_t^w is nonincreasing with respect to t, and the last step is the conclusion of Corollary 4; this completes the proof. ☐

2.3. Proof of Theorem 1.

Proof. We discuss three cases: (1) the simplest case where all the hypotheses are null, and the null p-values are uniformly distributed; (2) the case where non-nulls may exist, and the null p-values are uniformly distributed; and finally (3) the case where non-nulls may exist, and the null p-values can be mirror-conservative.

Case 1: nulls only and null p-values uniform. By Lemma 1, $\{\mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right)\}_{j=1}^n$ are i.i.d. Bernoulli with parameter p_*. Observe that the stopping rule in Algorithm 1, $\widehat{\mathrm{FWER}}_t \equiv 1 - (1 - p_*)^{|\mathcal{R}_t^-| + 1} \le \alpha$, can be rewritten as $|\mathcal{R}_t^-| + 1 \le v$ where

$$v = \left\lfloor \frac{\log(1 - \alpha)}{\log(1 - p_*)} \right\rfloor, \tag{14}$$

which is also equivalent as $|\mathcal{R}_t^-| < v$. We show that the number of false rejections is stochastically dominated by $\mathrm{NB}(v, p_*)$ by Corollary 3. Let $A_j = \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right)$ and $\widetilde{M}_t = \sum_{j=1}^{n-t} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right)$. The stopping time is $\widetilde{\tau} = \min\{0 \le t < n : |\mathcal{R}_t^-| = (n - t) - \widetilde{M}_t < v \text{ or } t = n - 1\}$. The number of rejections at the stopping time is

$$|\mathcal{R}_{\widetilde{\tau}}^+| \equiv \sum_{j=1}^{n-\widetilde{\tau}} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right) \equiv \widetilde{M_{\widetilde{\tau}}} \preceq \mathrm{NB}(v, p_*),$$

where the last step is the conclusion of Corollary 3. Note that we assume all the hypotheses are null, so the number of false rejections is $|\mathcal{R}_{\tilde{\tau}}^+ \cap \mathcal{H}_0| = |\mathcal{R}_{\tilde{\tau}}^+| \preceq \mathrm{NB}(v, p_*)$. Thus, FWER is upper bounded:

$$\mathbb{P}(|\mathcal{R}_{\tilde{\tau}}^+ \cap \mathcal{H}_0| \geq 1) \leq 1 - (1-p_*)^v \leq \alpha, \tag{15}$$

where the last inequality follows by the definition of v in (14). Thus, we have proved FWER control in Case 1.

Remark: This argument also provides some intuition on the FWER estimator (4): $\widehat{\mathrm{FWER}}_t = 1-(1-p_*)^{|\mathcal{R}_t^-|+1}$. Imagine we run the algorithm for one time without any stopping rule until time t_0 to get an instance of $\widehat{\mathrm{FWER}}_{t_0}$, then we run the algorithm on another independent dataset, which stops once $\widehat{\mathrm{FWER}}_t \leq \widehat{\mathrm{FWER}}_{t_0}$. Then in the second run, FWER is controlled at level $\widehat{\mathrm{FWER}}_{t_0}$.

Case 2: non-nulls may exist and null p-values are uniform. We again argue that the number of false rejections is stochastically dominated by $\mathrm{NB}(v, p_*)$, and in this case we use Corollary 4. Consider $A_j = \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right)$ and $W_j = \mathbb{1}\left(\pi_j \in \mathcal{H}_0\right)$, which satisfies condition (b) in Corollary 4 according to Corollary 1. Let $m = |\mathcal{H}_0|$, then $\sum_{j=1}^n W_j = m$, which corresponds to condition (a). Imagine an algorithm stops once

$$\sum_{j=1}^{n-t} \mathbb{1}\left(h\left(P_{\pi_j}\right) = -1 \cap \pi_j \in \mathcal{H}_0\right) = \sum_{j=1}^{n-t} W_j(1-A_j) < v, \tag{16}$$

and we denote the stopping time as τ^w. By Corollary 4, the number of false rejections in this imaginary case is

$$\sum_{j=1}^{n-\tau^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1 \cap \pi_j \in \mathcal{H}_0\right) = \sum_{j=1}^{n-t} W_j A_j = M_{\tau^w}^w \preceq \mathrm{NB}(v, p_*).$$

Now, consider the actual i-FWER test which stops when $|R_t^-| = (n-t) - \sum_{j=1}^{n-t} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right) < v$, and denote the true stopping time as τ_T^w. Notice that at the stopping time, it holds that

$$\begin{aligned} &\sum_{j=1}^{n-\tau_T^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = -1 \cap \pi_j \in \mathcal{H}_0\right) \\ \leq\ &\sum_{j=1}^{n-\tau_T^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = -1\right) \\ =\ &(n-\tau_T^w) - \sum_{j=1}^{n-\tau_T^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right) < v, \end{aligned}$$

which means that stopping rule (16) is satisfied at τ_T^w. Thus, $\tau_T^w \geq \tau^w$ and $M_{\tau_T^w}^w \leq M_{\tau^w}^w$ (because M_t^w is nonincreasing with respect to t). It follows that the number of false rejections is

$$|\mathcal{R}_{\tau_C^w}^+ \cap \mathcal{H}_0| \equiv \sum_{j=1}^{n-\tau_C^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1 \cap \pi_j \in \mathcal{H}_0\right) \equiv M_{\tau_C^w}^w \leq M_{\tau^w}^w \preceq \mathrm{NB}(v, p_*).$$

We then prove FWER control using a similar argument as (15):

$$\mathbb{P}(|\mathcal{R}_{\tau^w}^+ \cap \mathcal{H}_0| \geq 1) \leq 1 - (1-p_*)^v \leq \alpha,$$

which completes the proof of Case 2.

Case 3: non-nulls may exist and null p-values can be mirror-conservative. In this case, we follow the proof of Case 2 except additionally conditioning on all the masked p-values, $\{g(P_{\pi_k})\}_{k=1}^n$. By Corollary 2 and Corollary 5, we again conclude that the number of false rejections is dominated by a negative binomial:

$$|\mathcal{R}_{\tau_C^w}^+ \cap \mathcal{H}_0| \preceq \mathrm{NB}(v, p_*),$$

if given $\{g(P_{\pi_k})\}_{k=1}^n$. Thus, FWER conditional on $\{g(P_{\pi_k})\}_{k=1}^n$ is upper bounded:

$$\mathbb{P}\left(|\mathcal{R}_{\tau^w}^+ \cap \mathcal{H}_0| \geq 1 \middle| \{g(P_{\pi_k})\}_{k=1}^n\right) \leq 1 - (1-p_*)^v \leq \alpha,$$

which implies the FWER control by the law of iterated expectations. This completes the proof of Theorem 1. □

3. An alternative perspective: closed testing

This section summarizes the comments from Jelle Goeman, who kindly points out the connection between our proposed method and the *closed testing* (Marcus et al., 1976). Closed testing is a general framework that generates a procedure with FWER control given any test with Type 1 error control. Specifically, we reject H_i if all possible sets of hypotheses involving H_i, denoted as $U \ni i$, can be rejected by a "local" test for hypotheses in U with Type 1 error control at level α.

The i-FWER test we propose shares some commonalities with the *fallback procedure* (Wiens & Dmitrienko, 2005), which can be viewed as a shortcut of a closed testing procedure. We briefly describe the commonalities and differences next. Let v be a prespecified positive integer. The fallback procedure orders the hypotheses from most to least interesting, and proceeds to test them one by one at level α/v until it has failed to reject v hypotheses. The hypothesis ordering is allowed to be data-dependent as long as the ordering is independent of the p-values, corresponding to ordering by the side information x_i in our language. This procedure is essentially also what the i-FWER test does except (a) the i-FWER test uses the Šidák correction instead of the Bonferroni correction; (b) we are interested in whether rejecting each hypothesis instead of adjusting individual p-values, so the ordering only needs to be independent of reject/non-reject status instead of on the full p-values, which allows us to split each p-value into $h(P_i)$ and $g(P_i)$; (c) under the assumption of independent null p-values, we are allowed to use the p-values excluded from the candidate rejection set $\mathcal{R}_t$ as independent information to create the ordering. The latter two differences enable the i-FWER test to be interactive based on a considerably large amount of data information.

3.1. Alternative proof of Theorem 1

The above observation leads to a simple proof of the error control guarantee without involving any martingales or negative binomial distributions, once we rewrite the i-FWER test in the language of closed testing.

Proof. For simplicity, we consider the nulls with only uniform p-values. Let v be a prespecified positive integer, and define $p_* = 1 - (1-\alpha)^{1/v}$. Imagine that the i-FWER test does not have a stopping rule and let $\pi_n, \ldots, \pi_1$ be the order in which the hypotheses are chosen by an analyst, where each choice π_t can base on all the information in $\mathcal{F}_{n-t}$.

Here, we construct a closed testing procedure by defining a local test with Type 1 error control for an arbitrary subset $U \in [n]$ of size $|U|$. Sort the hypotheses in U according to the analyst-specified ordering from the last π_n to the first chosen π_1. If the number of hypotheses in U is larger than v, define U_v as the subset of U of size v corresponding to the hypotheses in U that are chosen last. For example, if $U = [n]$, we have $U_v = \{\pi_v, \ldots, \pi_1\}$. If $|U| \leq v$, define $U_v = U$. We reject the subset U if $h(P_i) = 1$ (i.e., $P_i \leq p_*$) for at least one $i \in U_v$. This is a valid local test, since it controls the Type 1 error when all the hypotheses in U are null. To verify the error control, notice that $h(P_i)$'s are independent and follows Bernoulli(p_*), and U_v is independent of $\{h(P_i)\}_{i \in U_v}$ by the construction of sequence $\pi_1, ..., \pi_n$, so the Type 1 error satisfies

$$\mathbb{P}(\exists i \in U_v : h(P_i) = 1) \leq 1 - (1 - p_*)^v,$$

which is less than α by the definition of v and p_*. Indeed, the local test corresponds to a Šidák correction for v number of hypotheses. Through closed testing, this local test leads to a valid test with FWER control.

Next, we show that the rejection set from the i-FWER test, $\mathcal{R}_\tau^+$, is included in the rejection set from the above closed testing procedure. Choose any hypothesis $j \in \mathcal{R}_\tau^+$ and any set $W \ni j$. If H_j is among the last v hypotheses last chosen in W (or if $|W| \leq v$), the local test for W reject the null since $P_j \leq p_*$ by the definition of $\mathcal{R}_\tau^+$. Otherwise, the v hypotheses last chosen in W are all chosen after H_j. Since $j \in \mathcal{R}_\tau^+$ and by the definition of τ, we have $|\mathcal{R}_\tau^-| \leq v - 1$. That is, there can be at most $v - 1$ hypotheses among these v such that $h(P_i) = -1$, so set W is rejected by the local test as described in the previous paragraph. It follows from the definition of FWER and the error control of the larger (or equivalent) rejection set from the closed testing procedure that $\mathcal{R}_\tau^+$ has FWER control. $\square$

3.2. Improvement on an edge case

From the closed testing procedure constructed in the above proof, we observe that the local tests do not exhaust the α-level for intersections of less than v hypotheses. This suboptimality can be remedied, but it will only improve power for rejecting all hypotheses given that almost all are already rejected (i.e., most subsets U with $|U| > v$ are rejected by the local test). In the i-FWER test, such a case potentially corresponds to the case where the initial rejection set has less than v hypotheses with negative $h(P_i)$, so the algorithm stops before shrinking $\mathcal{R}_0$, and reject all the hypotheses with positive $h(P_i)$. However, we might not fully use the error budget because $\widehat{\text{FWER}}_0 < \alpha$. However, we might not fully use the error budget because $\widehat{\text{FWER}}_0 < \alpha$. To improve power and efficiently use all the error budget, we propose randomly rejecting the hypotheses with a negative $h(P_i)$ if the algorithm stops at step 0.

Recall that the number of negative $h(P_i)$ is $|\mathcal{R}_0^-|$. For each hypothesis with a negative $h(P_i)$, we independently decide to reject it with probability $1-(1-\alpha_{\text{re}})^{1/|\mathcal{R}_0^-|}$, where $\alpha_{\text{re}} := \alpha - \widehat{\text{FWER}}_0$ denotes the remaining error budget after rejecting all the hypotheses with positive $h(P_i)$'s. To see the error control guarantee of this improved algorithm, notice that

$$
\begin{aligned}
& \mathbb{P}(\exists i \in \mathcal{H}_0 : H_i \text{ is rejected}) \\
\leq\ & \mathbb{P}(\exists i \in \mathcal{H}_0 : h(P_i) = 1) + \mathbb{P}(\exists i \in \mathcal{H}_0 : h(P_i) = -1 \text{ and } H_i \text{ is rejected}) \\
\leq\ & \widehat{\text{FWER}}_0 + \mathbb{P}(\exists i \in \mathcal{R}_0^- : H_i \text{ is rejected}) \\
\leq\ & \widehat{\text{FWER}}_0 + \alpha_{\text{re}} = \alpha,
\end{aligned}
$$

where $\mathbb{P}(\exists i \in \mathcal{H}_0 : h(P_i) = 1) \leq \widehat{\text{FWER}}_0$ follows the argument using negative binomial distribution as in the proof of the original algorithm; and $\mathbb{P}(\exists i \in \mathcal{R}_0^- : H_i \text{ is rejected}) \leq \alpha_{\text{re}}$ is the result of a Šidák correction. We summarize the adjusted i-FWER test in Algorithm 1.

Algorithm 1 The adjusted i-FWER test

Input: Side information and p-values $\{x_i, P_i\}_{i=1}^n$, target FWER level α, and parameter p_*;
Procedure:
Initialize $\mathcal{R}_0 = [n]$;
if $\widehat{\text{FWER}}_0 \equiv 1-(1-p_*)^{|\mathcal{R}_0^-|+1} \leq \alpha$ **then**
  Obtain n independent indicators from a Bernoulli distribution with probability $1-(1-\alpha+\widehat{\text{FWER}}_0)^{1/|\mathcal{R}_0^-|}$, denoted as $\{I_i\}_{i\in[n]}$;
  Reject $\{H_i : i \in [n], h(P_i) = 1 \text{ or } I_i = 1\}$ and exit;
else
  for $t = 1$ **to** n **do**
    1. Pick any $i_t^* \in \mathcal{R}_{t-1}$, using $\{x_i, g(P_i)\}_{i=1}^n$ and $\{h(P_i)\}_{i\notin\mathcal{R}_{t-1}}$;
    2. Exclude i_t^* and update $\mathcal{R}_t = \mathcal{R}_{t-1}\backslash\{i_t^*\}$;
    if $\widehat{\text{FWER}}_t \equiv 1-(1-p_*)^{|\mathcal{R}_t^-|+1} \leq \alpha$ **then**
      Reject $\{H_i : i \in \mathcal{R}_t, h(P_i) = 1\}$ and exit;
    end if
  end for
end if

4. Sensitivity analysis

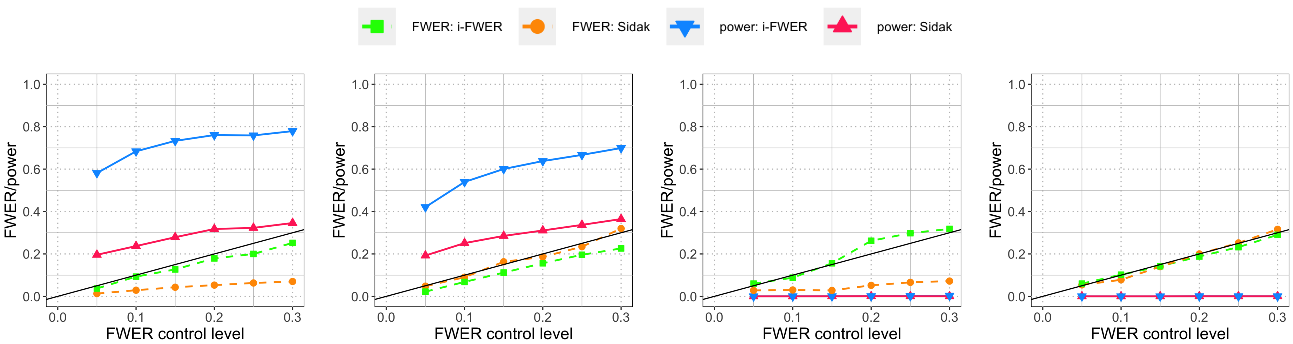
The i-FWER test is proved to have valid error control when the nulls are mutually independent and independent of the non-nulls. In this section, we evaluate the performance of the i-FWER test under correlated p-values. Our numerical experiments construct a grid of hypotheses as described in the setting in Section 3.2. The p-values are generated as

$$P_i = 1 - \Phi(Z_i), \text{where } Z = (Z_1, \ldots, Z_n) \sim N(\mu, \Sigma), \tag{17}$$

where $\mu = 0$ for the nulls and $\mu = 3$ for the non-nulls. The covariance matrix Σ, which is identity matrix in the main paper, is now set to an equi-correlated matrix:

$$\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}. \tag{18}$$

Under both the positively correlated case ($\rho = 0.5$) and the negatively correlated case ($-\rho = 0.5/n$ to guarantee that Σ is positive semi-definite), the i-FWER test seems to maintain the FWER control at most target levels even when all the hypotheses are nulls (see Figure 1c and 1d), and has higher power than the Šidák correction (see Figure 1a and 1b).



(a) Positively correlated case where $\rho = 0.5$ in the covariance matrix (18). The non-null mean value is 3.

(b) Negatively correlated case where $\rho = -0.5/n$ in the covariance matrix (18). The non-null mean value is 3.

(c) Negatively correlated case where $\rho = -0.5/n$ in the covariance matrix (18). All hypotheses are nulls.

(d) Negatively correlated case where $\rho = -0.5/n$ in the covariance matrix (18). All hypotheses are nulls.

Figure 1. FWER and power of the i-FWER test and the Šidák correction for dependent p-values generated by Gaussians as in (17) with covariance matrix (18) when the targeted level of FWER control varies in $(0.05, 0.1, 0.15, 0.2, 0.25, 0.3)$. The i-FWER test appears to control FWER below the targeted level and has relatively high power.

5. More results on the application to genetic data

Section 5 presents the number of rejections of the i-FWER test when the masking uses the tent function. We evaluate the i-FWER test when using the other three masking functions under the same experiments, but for simplicity, we only present the result when the FWER control is at level $\alpha = 0.2$ (see Table 1). Overall, the gap function leads to a similar number of rejections as the tent function, consistent with the numerical experiments. However, the railway (gap-railway) function leads to fewer rejections than the tent (gap) function, which seems counterintuitive. Upon a closer look at the p-values, we find that the null p-values are not uniform or have an increasing density (see Figure 2). As a result, when using the tent function, there are fewer masked p-values from the nulls that could be confused with those of the non-nulls (with huge p-values), compared with using the railway function where the masked p-values of the confused nulls are those close to the masking parameter (around 0.02).

Table 1. Number of rejections by i-FWER test using different masking functions when $\alpha = 0.2$. The tent function and the gap function leads to more rejections compared with the railway function and the gap-railway function. The parameters in the gap and gap-railway function are set to $p_l = p_*$ and $p_u = 0.5$, and we need $p_l < \alpha/2$ for the test to make any rejection under level α.

Masing function	$p_* = \alpha/2$	$p_* = \alpha/10$	$p_* = \alpha/20$
Tent	1752	1848	1794
Railway	1778	1463	1425
Gap	NA	1802	1846
Gap-railway	NA	1764	1788

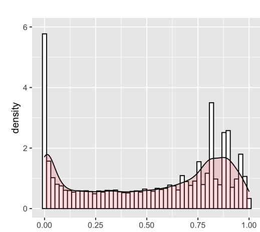


Figure 2. Histogram of p-values in the airway dataset. The number of p-values that are close to one is less than those that are close to the cutting point of the masking function (say 0.02). Consequently, the tent (gap) function leads to more rejections than the railway (gap-railway) function.

6. Error control for other masking functions

The proof in Appendix 2 is for the i-FWER test with the original tent masking function. In this section, we check the error control for two new masking functions introduced in Section 4.

6.1. The railway function

We show that the i-FWER test with the "railway" function (7) has FWER control, if the null p-values have convex CDF or nondecreasing f (recall f is the probability mass function for discrete p-values or the density function otherwise). We again assume the same independence structure as in Theorem 1 that the null p-values are mutually independent and independent of the non-nulls.

The proof in Appendix 2 implies that under the same independence assumption, the FWER control is valid if the null p-values satisfy condition (4). When using the railway masking function, condition (4) is indeed satisfied if the null p-values have nondecreasing f since

$$\begin{aligned}\mathbb{P}(h(P) = 1 \mid g(P) = a) = &\ \frac{p_* f(a)}{p_* f(a) + (1-p_*) f(\frac{1-p_*}{p_*} a + p_*)} \\ = &\ \frac{p_*}{p_* + (1-p_*) f(\frac{1-p_*}{p_*} a + p_*)/f(a)} \\ \leq &\ p_*,\end{aligned}$$

for every $a \in (0, p_*)$. Then, we can prove the FWER control following the same argument as Appendix 2.

6.2. The gap function

The essential difference of using the gap function instead of the tent function is that here, $\mathbb{1}\left(h(P) = 1\right)$ for the nulls follow a Bernoulli distribution with a different parameter, $\widetilde{p} = \mathbb{P}(P = 1 \mid P < p_l \text{ or } P > p_u) = \frac{p_l}{p_l + 1 - p_u}$. Once replacing p_* by $\widetilde{p}$, we get a the new FWER estimator $\widehat{\text{FWER}}_t$ as defined in (9) and the error control can be proved following Appendix 2.

7. Varying the parameters in the presented masking functions

We first discuss the original tent masking (2), which represents a class of masking functions parameterized by p_*. Similar to the discussion in Section 4, varying p_* also changes the amount of p-value information distributed to $g(P)$ for interaction (to exclude possible nulls) and $h(P)$ for error control (by estimating FWER), potentially influencing the test performance. On one hand, the masking function with smaller p_* effectively distributes less information to $g(P)$, in that a larger range of big p-values is mapped to small $g(P)$ (see Figure 7a). In such a case, the true non-nulls with small p-values and small $g(P)$ are less distinctive, making it difficult to exclude the nulls from $\mathcal{R}_t$. On the other hand, the rejected hypotheses in $\mathcal{R}_t^+$ must satisfy $P < p_*$, so smaller p_* leads to less false rejections given the same $\mathcal{R}_t$.

Experiments show little change in power when varying the value of p_* in $(0, \alpha)$ as long as it is not near zero, as it would leave little information in $g(P)$. Our simulations follow the setting in Section 3.2, where the alternative mean value is fixed at $\mu = 3$. We tried seven values of p_* as $(0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2)$, and the power of the i-FWER test does not change much for $p_* \in (0.05, 0.2)$. This trend also holds when varying the mean value of non-nulls, the size of the grid (with a fixed number of non-nulls), and the number of non-nulls (with a fixed size of the grid). In general, the choice of p_* does not have much influence on the power, and a default choice can be $p_* = \alpha/2$.

There are also parameters in two other masking functions proposed in Section 4. The railway function flips the tent function without changing the distribution of p-value information, hence the effect of varying p_* should be similar to the case in the tent function. The gap function (8) has two parameters: p_l and p_u. The tradeoff between information for interaction and error control exhibits in both values of p_l and p_u: as p_l decreases (or p_u increases), more p-values are available to the analyst from the start, guiding the procedure of shrinking $\mathcal{R}_t$, while the estimation of FWER becomes less accurate. Whether revealing more information for interaction should depend on the problem settings, such as the amount of prior knowledge.

8. Mixture model for the non-null likelihoods

Two groups model for the p-values. Define the Z-score for hypothesis H_i as $Z_i = \Phi^{-1}(1-P_i)$, where Φ^{-1} is the inverse function of the CDF of a standard Gaussian. Instead of modeling the p-values, we choose to model the Z-scores since when testing the mean of Gaussian as in (1), Z-scores are distributed as a Gaussian either under the null or the alternative:

$$H_0 : Z_i \overset{d}{=} N(0,1) \quad \text{versus} \quad H_1 : Z_i \overset{d}{=} N(\mu,1),$$

where μ is the mean value for all the non-nulls. We model Z_i by a mixture of Gaussians:

$$Z_i \overset{d}{=} (1-q_i)N(0,1) + q_i N(\mu,1), \text{ with } q_i \overset{d}{=} \text{Bernoulli}(\pi_i),$$

where q_i is the indicator of whether the hypothesis H_i is truly non-null.

The non-null structures are imposed by the constraints on π_i, the probability of being non-null. In our examples, the blocked non-null structure is encoded by fitting π_i as a smooth function of the hypothesis position (coordinates) x_i, specifically as a logistic regression model on a spline basis $B(x) = (B_1(x), \dots, B_m(x))$:

$$\pi_\beta(x_i) = \frac{1}{1+\exp(-\beta^T B(x_i))}, \tag{19}$$

EM framework to estimate the non-null likelihoods. An EM algorithm is used to train the model. Specifically we treat the p-values as the hidden variables, and the masked p-values $g(P)$ as observed. In terms of the Z-scores, Z_i is a hidden variable and the observed variable $\tilde{Z}_i$ is

$$\tilde{Z}_i = \begin{cases} Z_i, & \text{if } Z_i > \Phi^{-1}(1-p_*), \\ t(Z_i), & \text{otherwise,} \end{cases}$$

where $t(Z_i)$ depends on the form of masking. The updates needs values of its inverse function $t^{-1}(\tilde{Z}_i)$ and the derivative of $t^{-1}(\cdot)$, denoted as $\left(t^{-1}\right)'(\tilde{Z}_i)$, whose exact forms are presented below.

1. For tent masking (2),

$$\begin{aligned} t(Z_i) =&\ \Phi^{-1}\left[1 - \frac{p_*}{1-p_*}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) =&\ \Phi^{-1}\left[\frac{1-p_*}{p_*}\left(1-\Phi(\tilde{Z}_i)\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) =&\ -\frac{1-p_*}{p_*}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right), \end{aligned}$$

   where $\phi(\cdot)$ is the density function of standard Gaussian.

2. For railway masking (7),

$$\begin{aligned} t(Z_i) =&\ \Phi^{-1}\left[1 - p_* + \frac{p_*}{1-p_*}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) =&\ \Phi^{-1}\left[\frac{1-p_*}{p_*}\left(\Phi(\tilde{Z}_i) - 1 + p_*\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) =&\ \frac{1-p_*}{p_*}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right). \end{aligned}$$

3. For gap masking (8),

$$\begin{aligned} t(Z_i) =&\ \Phi^{-1}\left[1 - \frac{p_l}{1-p^u}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) =&\ \Phi^{-1}\left[\frac{1-p_u}{p_l}\left(1-\Phi(\tilde{Z}_i)\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) =&\ -\frac{1-p_u}{p_l}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right). \end{aligned}$$

if $Z_i < \Phi^{-1}(1-p_u)$. If $\Phi^{-1}(1-p_u) \le Z_i \le \Phi^{-1}(1-p_l)$, which corresponds to the skipped p-value between p_l and p_u, then $\tilde{Z}_i = Z_i$.

4. For gap-railway masking (10),

$$\begin{aligned} t(Z_i) =&\ \Phi^{-1}\left[1 - \frac{p_l}{1-p^u}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) =&\ \Phi^{-1}\left[\frac{1-p_u}{p_l}\left(\Phi(\tilde{Z}_i) - 1 + p_l\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) =&\ \frac{1-p_u}{p_l}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right). \end{aligned}$$

if $Z_i < \Phi^{-1}(1-p_u)$. If $\Phi^{-1}(1-p_u) \le Z_i \le \Phi^{-1}(1-p_l)$, which corresponds to the skipped p-value between p_l and p_u, then $\tilde{Z}_i = Z_i$.

Define two sequences of hypothetical labels $w_i = \mathbb{1}\{Z_i = \tilde{Z}_i\}$ and $q_i = \mathbb{1}\{H_i = 1\}$, where $H_i = 1$ means hypothesis i is truly non-null ($H_i = 0$ otherwise). The log-likelihood of observing $\tilde{Z}_i$ is

$$\begin{aligned} l(\tilde{Z}_i) =&\ w_i q_i \log\left\{\pi_i\phi\left(\tilde{Z}_i - \mu\right)\right\} + w_i(1-q_i)\log\left\{(1-\pi_i)\phi\left(\tilde{Z}_i\right)\right\} \\ &+ (1-w_i)q_i\log\left\{\pi_i\phi\left(t^{-1}(\tilde{Z}_i) - \mu\right)\right\} + (1-w_i)(1-q_i)\log\left\{(1-\pi_i)\phi\left(t^{-1}(\tilde{Z}_i)\right)\right\}. \end{aligned}$$

The E-step updates w_i, q_i. Notice that w_i and q_i are not independent, and hence we update the joint distribution of (w_i, q_i), namely

$$\mathbb{E}[w_i q_i] =: a_i, \quad \mathbb{E}[w_i(1-q_i)] =: b_i, \quad \mathbb{E}[(1-w_i)q_i] =: c_i, \quad \mathbb{E}[(1-w_i)(1-q_i)] =: d_i,$$

where $a_i + b_i + c_i + d_i = 1$. To simplify the expression for updates, we denote

$$L_i := \pi_i\phi\left(\tilde{Z}_i - \mu\right) + (1-\pi_i)\phi\left(\tilde{Z}_i\right) + \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right|\pi_i\phi\left(t^{-1}(\tilde{Z}_i) - \mu\right) + \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right|(1-\pi_i)\phi\left(t^{-1}(\tilde{Z}_i)\right).$$

For the hypothesis i whose p-value is masked, the updates are

$$\begin{aligned} a_{i,\text{new}} =&\ \mathbb{E}[w_i q_i \mid \tilde{Z}_i] = \pi_i\phi\left(\tilde{Z}_i - \mu\right)\Big/L_i; \\ b_{i,\text{new}} =&\ \mathbb{E}[w_i(1-q_i) \mid \tilde{Z}_i] = (1-\pi_i)\phi\left(\tilde{Z}_i\right)\Big/L_i; \\ c_{i,\text{new}} =&\ \mathbb{E}[(1-w_i)q_i \mid \tilde{Z}_i] = \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right|\pi_i\phi\left(t^{-1}(\tilde{Z}_i) - \mu\right)\Big/L_i; \\ d_{i,\text{new}} =&\ \mathbb{E}[(1-w_i)(1-q_i) \mid \tilde{Z}_i] = \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right|(1-\pi_i)\phi\left(t^{-1}(\tilde{Z}_i)\right)\Big/L_i. \end{aligned}$$

If the p-value is unmasked for i, the updates are

$$\begin{aligned} a_{i,\text{new}} =&\ \left(1 + \frac{(1-\pi_i)\phi\left(\tilde{Z}_i\right)}{\pi_i\phi\left(\tilde{Z}_i - \mu\right)}\right)^{-1}; \\ b_{i,\text{new}} =&\ 1 - a_{i,\text{new}}; \quad c_{i,\text{new}} = 0; \quad d_{i,\text{new}} = 0. \end{aligned}$$

In the M-step, parameters μ and β (in model (19) for π_i) are updated. The update for μ is

$$\mu_{\text{new}} = \operatorname*{argmax}_{\mu}\sum_i l(\tilde{Z}_i) = \frac{\sum a_i\tilde{Z}_i + c_i t^{-1}(\tilde{Z}_i)}{\sum a_i + c_i}.$$

The update for β is

$$\beta_{\text{new}} = \operatorname*{argmax}_{\beta}\sum_i (a_i + c_i)\log\pi_\beta(x_i) + (1 - a_i - c_i)\log(1 - \pi_\beta(x_i)),$$

where $\pi_\beta(x_i)$ is defined in equation (19). It is equivalent to the solution of GLM (generalized linear model) with the logit link function on data $\{a_i + c_i\}$ using covariates $\{B(x_i)\}$.

References

Lei, L. and Fithian, W. AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679, 2018.

Marcus, R., Eric, P., and Gabriel, K. R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Wiens, B. L. and Dmitrienko, A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*, 15(6):929–942, 2005.

Zhao, Q., Small, D. S., and Su, W. Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association*, 114(527): 1291–1304, 2019.